doi:10.1111/j.1420-9101.2006.01161.x

COMMENTARY

Various remarks on Lehmann and Keller's article

A. GRAFEN

Zoology Department, Oxford University, Oxford, UK

The fundamental message of Lehmann & Keller is absolutely right, namely that a large literature on altruism and cooperation has failed to grasp the significance and power of the original analysis and classification of Hamilton (1964, 1970), and that authors should be encouraged to interpret their conclusions within a Hamiltonian framework rather than claim an often spurious novelty at a fundamental level. A generation ago, I published in the same genre (Grafen, 1985) in relation to then prevalent strands of literature on altruism and cooperation. More recently, I have contributed with others (Axelrod *et al.*, 2004) to the current literature on altruism and cooperation in viscous populations with explicit spatial structure – and offered precisely an inclusive fitness interpretation of the results of the simulations and an analytical approach that would facilitate such an interpretation more widely.

The value of situating work in relation to a Hamiltonian framework stems from the biologically fundamental significance of Hamilton's analysis, as well as from the conceptual integration of a body of work that derives from having a single central theory of reference.

Let me begin by discussing the foundational work available on inclusive fitness. I have discussed the history in more detail elsewhere (Grafen, 2003, 2004): in brief, Hamilton's original work was extraordinarily influential in an informal way, but from a technical point of view was more or less ignored. Later derivations of inclusive fitness started from scratch and not from Hamilton's own models. There were two reasons for this. Hamilton's work had certain technical difficulties that made it difficult for population geneticists of the time to take up and develop. But the main problem was that Hamilton claimed to be establishing an optimization principle in a complicated situation with social interactions, just at the time when population geneticists were rejecting the optimization principle Fisher (1930) had proposed in the simpler, non-social case. Now that Fisher's fundamental theorem is properly understood (Price, 1972) and its truth accepted (Ewens, 1992; Edwards, 1994), it is time for a re-evaluation of Hamilton's work, and in a recently published paper (Grafen, 2006) I have produced a modern and fully explicit derivation of inclusive fitness, following very much in the methodological footsteps of Hamilton (1964, 1970), but adding a fully explicit account of what is meant by the optimization of fitness.

My paper in broad terms supports fully the claims of Hamilton (1964, 1970), and it is worth reiterating them. Unlike almost all the other work on inclusive fitness, Hamilton's original papers offer an extremely general model. The assumption in classical population genetics that had to be relaxed to permit the study of social behaviour was that the fertility or survivorship of an individual depended only on its own genotype. Hamilton's models permitted the fertility of one individual to depend in quite a general way on the genotypes of all the other individuals in the population, as well as her own. He assumed additivity of the effects, but justified this in terms of least squares regression as an approximation under weak selection. With this very broad assumption, he constructed a quantity he called inclusive fitness and showed that it played the role in the new model that classical fitness played in the non-social model, i.e. alleles associated with a higher 'fitness' increased in frequency.

If we accept Hamilton's argument, therefore, social behaviour in general must follow his analysis, as he has made very few assumptions. Virtually all models since have fallen within the ambit of the original model and have made very special assumptions within it. Most models assume that all social actions involve just two individuals, and furthermore that all interacting pairs play the same game; others assume a grouped population, and that all the members of an individual's group are equal recipients of her altruism and also that all acts of altruism have the same benefits and costs. These special cases were often analysed because Hamilton's analysis was thought inadequate, but usually it had already encompassed them from the beginning (Grafen, 1985).

In the literature reviewed by Lehmann and Keller, Hamilton's assumptions are, by and large, met too. Although Hamilton (1964) assumed random mating, the Price equation used by Hamilton (1970) allowed that assumption to be avoided, and so the spatial structure of some of the models does not contradict the basic assumptions of inclusive fitness. Non-additivity may require the assumption of weak selection and so restrict the absolute correctness of the inclusive fitness analysis to a linearized version close to some equilibrium. But this applies equally to Lehmann and Keller's own model, based on the Taylor–Frank approach (Taylor & Frank, 1996; Frank, 1998), and is quite adequate for an equilibrium analysis.

Lehmann and Keller's conclusion is therefore fully endorsed that the models reviewed fall within the basic structure of Hamilton's inclusive fitness and that they would be well interpreted in those terms.

After making a basically supportive point at a fundamental level, I now move on to engage more fully with some specific points in Lehmann and Keller's argument. Let me first note that Lehmann and Keller rightly recognize that the '*b*' and '*c*' to be used in Hamilton's

Correspondence: A. Grafen, St John's College, St Giles, Oxford, Oxon OX1 3JP, UK.
Tel.: 01865 277 438; fax: 01865 277 435;
e-mail: alan.grafen@sjc.ox.ac.uk

rule need to be appropriately derived. Inclusive fitness is a strong theory that dictates how its terms need to be measured – one cannot simply construct a model and assume that Hamilton's rule applies to parameters one has arbitrarily labelled 'r', 'b' and 'c', any more than a physicist would construct her own model of simple mechanics and expect Newton's second law '$F = ma$' to hold with arbitrarily labelled parameters 'F', 'm' and 'a'. The construction of the appropriate parameters can be seen explicitly in their equations (4) and (5), and is firmly based on the foundational work of Taylor & Frank (1996).

There is one level in the authors' model that seems unnecessarily complex to me, namely the one involving the parameter ζ. The model has B and C as the elemental parameters, which are assumed to be non-negative, so that the 'initial cost' is always positive and 'initial benefit' is always positive. A direct advantage to the actor would be a negative cost and is obtained by the device that a fraction ζ of the benefit returns directly to her, so that the 'net cost' is $C–\zeta B$. Further, the 'net benefit' to the recipient is reduced by the same amount to $(1–\zeta)B$. There is, however, nothing in the logic or algebra to prevent B and C themselves taking a negative sign, so they could instead be defined in the first place as the net benefit and the net cost, avoiding the need for ζ. I cannot myself see what biological value is gained by that extra level, which introduces an unwanted and restricting specificity. Cooperation and altruism do not seem to work that way: in general, there is no stage at which an actor has B offspring at her disposal and then has to decide how to allocate them, and can allocate them back to herself if she chooses. The practical mechanics of cooperation determine who will receive the benefits, at the same time as how large they will be, and the notional allocation parameterized by ζ seems wrong in the abstract. The same point can be put another way. There are social acts that have a negative B and/or C, and the whole theory applies equally in those cases with ζ set to zero. Why restrict the applicability of the theory to the minority special case where these negative values arise through some mechanism that is fairly represented by ζ? But perhaps I have missed something here. The extra level does assist in drawing some of the literature into the framework, but this is not a defence at a conceptual level: that part of the literature may simply be misconceived.

Lehmann and Keller note a potential discrepancy between simulation work on the evolutionary stability of continuing cooperation and the analytical result of Lorberbaum (1994) that the Repeated Prisoners' Dilemma has no evolutionarily stable strategy (ESS). There is a general conceptual point here that is worth developing. The Repeated Prisoners' Dilemma has a full strategy set and full rationality; i.e. we should imagine each player at each stage of the contest being able to consider the whole history of the contest, the likely response of her opponent to any course of action and the inferences the opponent may potentially draw from the player's own past and future behaviour. The proof that no ESS exists relies on this rich strategic situation. By contrast, the model of Lehmann and Keller is extremely reduced strategically. A player chooses two parameters, τ and β, representing the extent of initial cooperation and the slope relating the opponent's level of cooperation at each stage to the player's level of cooperation at the next stage, and these two parameters control the behaviour of the player through the whole sequence of stages. Thus, from a game theoretic point of view, this is not a sequential game at all, but instead a simple one-off game with a two-dimensional strategic choice for each player. (It would be a two-stage game if a player's choice of β were allowed to depend on her opponent's choice of τ, but I see no sign that is intended.) It is true that the form of the payoff function has been motivated by parallels with an iterated game, but none of that iteration survives into the strategic situation the players find themselves in. Let me make the immediate point that this shows there is no problem in reconciling the simulation and analytical results in the deterministic case, and also that the effect of errors will be very different in the two cases, and so Lehmann and Keller's concern is again unnecessary; and go on to make a more general remark. Game theory has contributed much to biology, and recently biology has been making contributions in return. These may have begun with Selten (1983) and the ideas involved are discussed by Samuelson (2002), one of the main protagonists. The now large literature may be accessed by searching for 'evolutionary game theory' online. Biology developed for itself and then inspired in other areas that use game theory a down-to-earth concrete approach to some of the more ethereal difficulties abstract game theory had found itself bewitched by. Biologists may well have taken their approach through a robust ignorance of the intellectual background, and I am certainly guilty of publishing on evolutionary games without understanding the distinction I draw attention to in this paragraph, but its influence in game theory has been no less real for that. Thus, there is a distinguished history of biologists productively confusing the 'purely rational' and the 'simple down-to-earth' approaches to game theory.

There are two points about recognition systems where I disagree with Lehmann and Keller's claims, and want to make the reasons plain here.

The first, rather minor, claim is that phenotype matching results in uniform genetic similarity over the whole genome. It is clear (for example from the 'telegraph wire' diagram in Grafen, 1985) that similarity will be high at loci that contribute to the matched phenotypic traits, and at linked loci, and low elsewhere. Clearly if the phenotype is matched on a set of traits that are affected by loci distributed throughout the genome, then the telegraph poles could in principle keep similarity high throughout the genome. But there are two problems here. One is that genetic similarity cannot be kept very high throughout the genome, as only a tiny fraction of individuals are similar enough, and the costs of searching would soon outweigh the benefits from obtaining a slightly higher similarity. The other is that I am unaware of any work that claims to show, even under idealized circumstances, that the telegraph poles would be at the same height throughout the genome, never mind the wires, which is what would be required to justify a claim of 'uniform genetic similarity'. Indeed at first thought it seems unlikely that a phenotype-matching mechanism would do that, unless the similarity were to be complete, and result in finding an actual or effectual identical twin. Note that matching on a quantitative trait will not achieve such twinning: matching could ensure similarity of the trait value alone, whereas the similarities at the individual contributing loci would be much weaker and depend on many incidental features. The claim for uniform genetic similarity does not derive from the authors' model and nor do they offer other arguments. It would be interesting to know what they had in mind. The interest of this point is that kinship is sometimes claimed (e.g. Grafen, 2006) to be the only biological factor that can produce uniform genetic similarity across the genome.

The second and more serious claim is that the linkage disequilibrium between altruist gene and recognition trait in any greenbeard system is bound to decay, with the result that the altruism will then disappear. There are sophisticated three-locus forces at play here, which the authors' model does not claim to capture, and they offer no other argument. Grafen (1990) discussed the forces verbally. Axelrod *et al.* (2004) show in an example that some kinds of biologically plausible situation can maintain an association between altruism and greenbeard-like recognition mechanisms. The key points are that 'tag alleles' are better indicators of relatedness when they are rare, setting up a negative frequency dependence at the tag-locus; and that provided there are enough alleles at that locus, and there is enough mixing up of the population to weed out free-riders by exposing them to individuals that do not share the same tag, all the alleles can be rare enough to support a greenbeard-like altruism. Now greenbeard models are notoriously subtle and full of traps for the unwary, and it may be that the authors had a different claim in mind to the one I have understood: but the statements in the target paper *seem* quite unequivocal.

A final issue is that Lehmann and Keller make two very specific claims in their abstract and in their paper. First, that altruism and cooperation can evolve as a result of a combination of only four elementary reasons; and second that there is sharp distinction between those four reasons on the one hand, which permit the evolution of cooperation, and coercion, punishment and policing on the other hand, which can only alter the threshold cost–benefit ratio. My difficulty is that they give the impression in the abstract that these claims are proved in the paper, but they do not seem to be. Lehmann and Keller make an admission in the first case by offering an inductive argument in their conclusion ('we are not aware of situations conducive to helping when at least one of our four conditions is not fulfilled') that would be superfluous if they had proved their point; and their model does not encompass coercion and so on, and thus is not capable of proving a delineation of the kind claimed. I have no reason to doubt either claim, but I do not regard them as established, but rather as interesting hypotheses. As made clear earlier, the main burden of the target article does not to my mind lie in these particular points.

To conclude, I see Lehmann and Keller's very positive contribution as capturing a range of important models on altruism and cooperation, many dealing with iterated games, within a synthetic model of their own, which they then interpret in Hamiltonian terms. This not only draws a large body of literature into the appropriate framework, but also sets the right example for future work.

References

Axelrod, R., Hammond, R.A. & Grafen, A. 2004. Altruism via kin-selection strategies that rely on arbitrary tags with which they coevolve. *Evolution* **58**: 1833–1838.

Edwards, A.W.F. 1994. The fundamental theorem of natural selection. *Biol. Rev.* **69**: 443–474.

Ewens, W.J. 1992. An optimizing principle of natural selection in evolutionary population genetics. *Theor. Popul. Biol.* **42**: 333–346.

Fisher, R.A. 1930. *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, UK.

Frank, S.A. 1998. *The Foundations of Social Evolution*. Princeton University Press, Princeton, NJ, USA.

Grafen, A. 1985. A geometric view of relatedness. *Oxf. Surv. Evol. Biol.* **2**: 28–89.

Grafen, A. 1990. Do animals really recognize kin? *Anim. Behav.* **39**: 42–54.

Grafen, A. 2003. Fisher the evolutionary biologist. *J. R. Stat. Soc. Ser. D (Stat.)* **52**: 319–329.

Grafen, A. 2004. William Donald Hamilton. *Biogr. Mem. Fellows R. Soc.* **50**: 109–132.

Grafen, A. 2006. Optimization of inclusive fitness. *J. Theor. Biol.* **238**, 541–563.

Hamilton, W.D. 1964. The genetical evolution of social behaviour. *J. Theor. Biol.* **7**, 1–52.

Hamilton, W.D. 1970. Selfish and spiteful behaviour in an evolutionary model. *Nature* **228**, 1218–1220.

Lehmann, L. & Keller, L. in press. The evolution of cooperation and altruism: a general framework and a classification of models. *J. Evol. Biol.* **19**: 1365–1376.

Lorberbaum, J. 1994. No strategy is evolutionarily stable in the repeated prisoners' dilemma. *J. Theor. Biol.* **168**: 117–130.

Price, G.R. 1972. Fisher's 'fundamental theorem' made clear. *Ann. Hum. Genet.* **36**: 129–140.

Samuelson, L. 2002. Evolution and game theory. *J. Econ. Perspect.* **16**: 47–66.

Selten, R. 1983. Evolutionary stability in extensive 2-person games. *Math. Soc. Sci.* **5**: 269–363.

Taylor, P.D. & Frank, S.A. 1996. How to make a kin selection model. *J. Theor. Biol.* **180**: 27–37.